THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Credit Default Prediction

Machine Learning Final Project

Group 5 : Sameer Batra & Jeongmin An

Table of Contents

Overview

- Project Goals
- Dataset overview

Exploratory Data Analysis

- Target Variable Analysis
- Age distribution
- Monthly Income Analysis
- Late payment Behavior
- Feature Correlation Analysis

Data Pre-processing

- Handling Missing Value
- Handling Outliers

Hyperparameter Tuning & Model selection

Project Goals

✅ Risk Assessment

- Predict future delinquency to strengthen risk monitoring
- Reduce financial losses by identifying high-risk borrowers early

✅ Support Data-Driven Credit Decisions

- Provide consistent and responsible lending decisions
- Deliver actionable insights into financial behaviors and risk drivers

✅ Build a reliable machine learning model

- Compare multiple models and select the best-performing approach
- Ensure strong performance through preprocessing, tuning, and evaluation

Dataset Overview

“Give Me Some Credit” Dataset

- ~150,000 records from Kaggle Competition
- The target variable, **SeriousDlqin2yrs,** indicates whether a person becomes 90+ days delinquent **within the two years,** making this a binary classification problem.

The dataset contains **11 variables,** including key financial indicators such as:

RevolvingUtilizationOfUnsecuredLines (credit usage rate)

DebtRatio

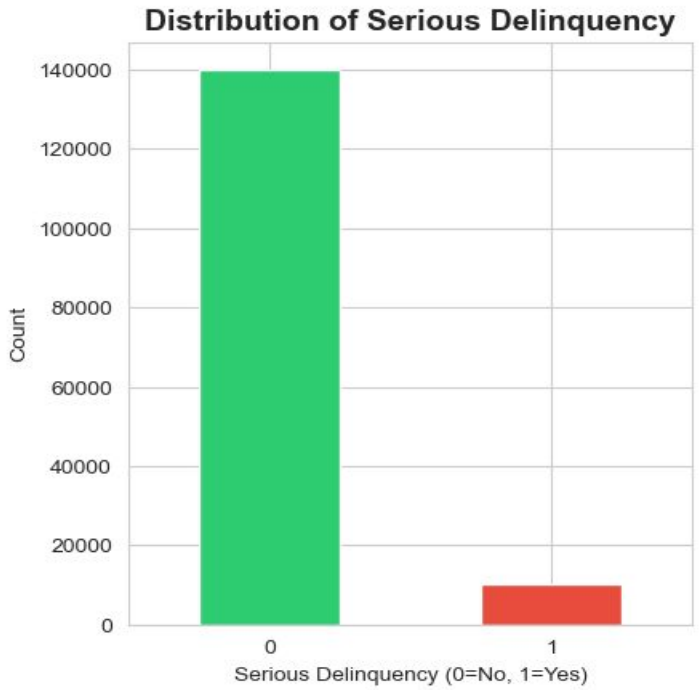
MonthlyIncome

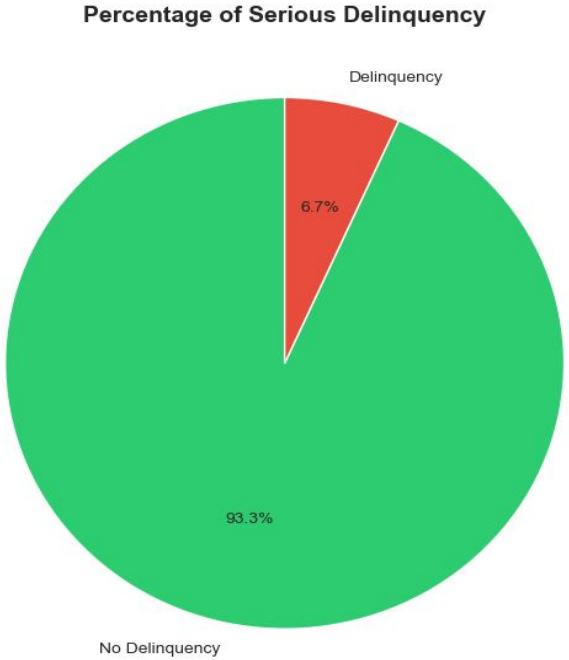
Delinquency history at 30–59, 60–89, and 90+ days

- **Number of open credit lines**
- **Real estate loans**
- **Dependents**

EDA

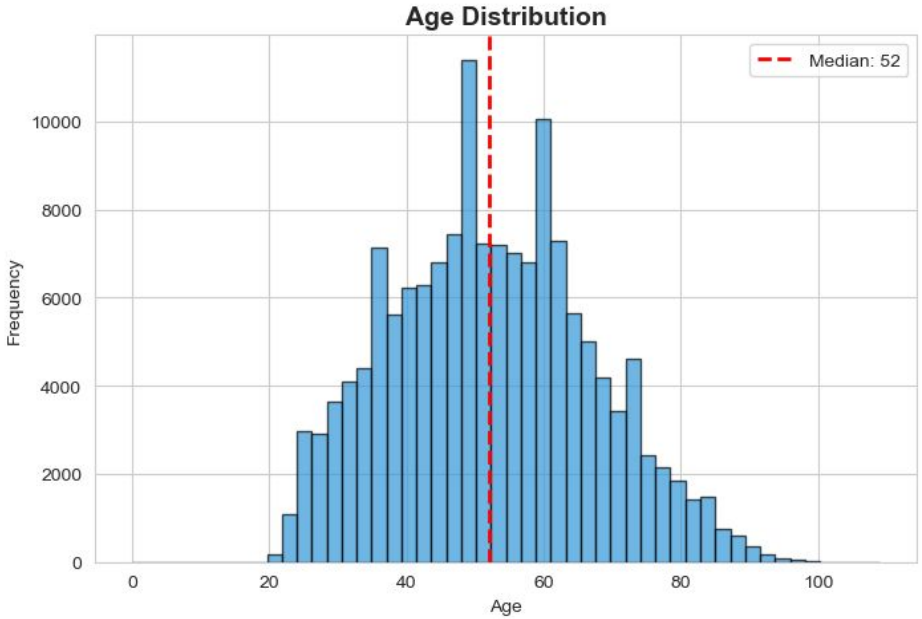
Target Variable Analysis

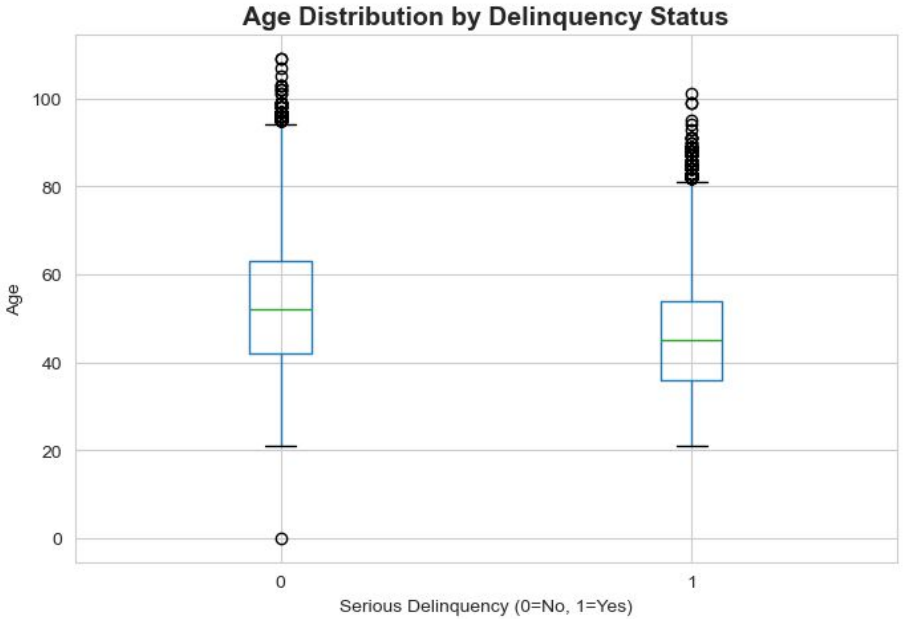




EDA

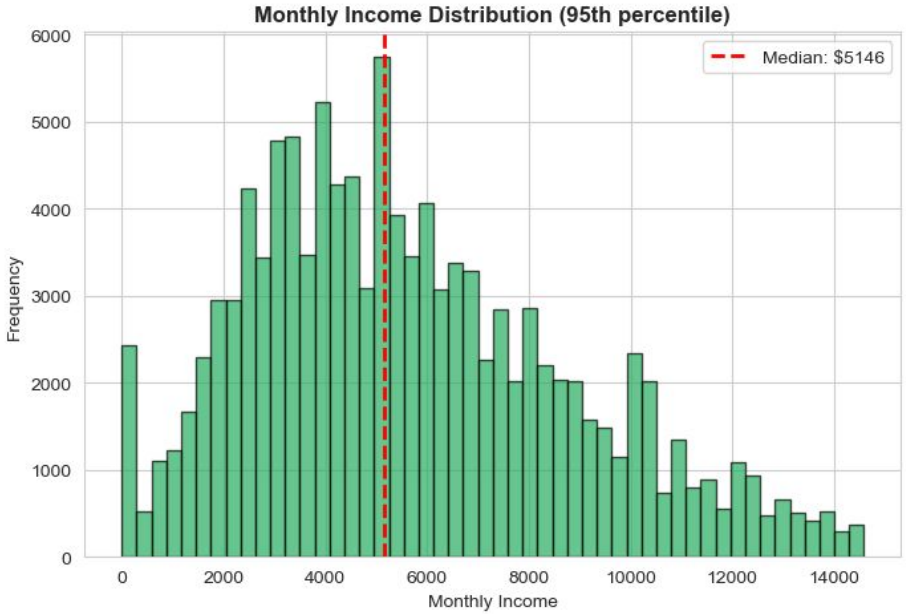
Age Distribution

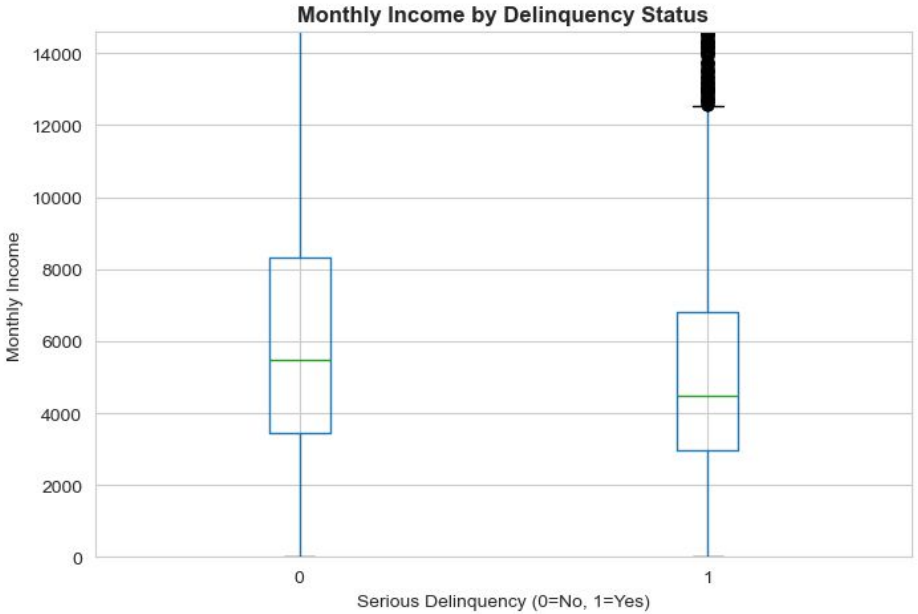




EDA

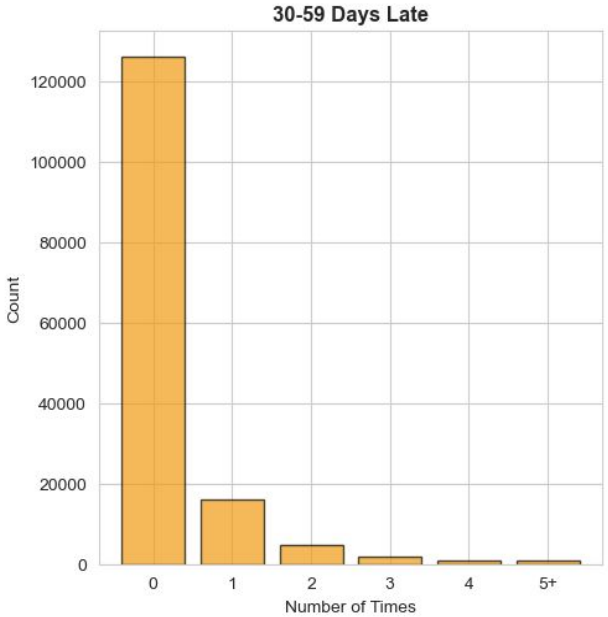
Monthly Income Analysis

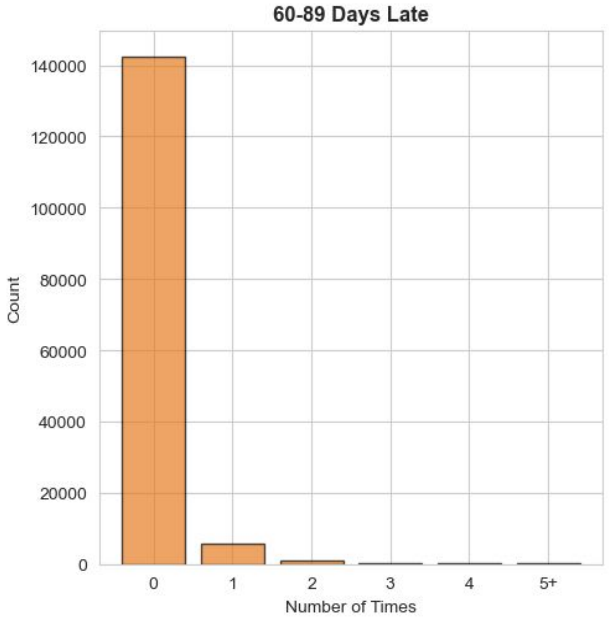


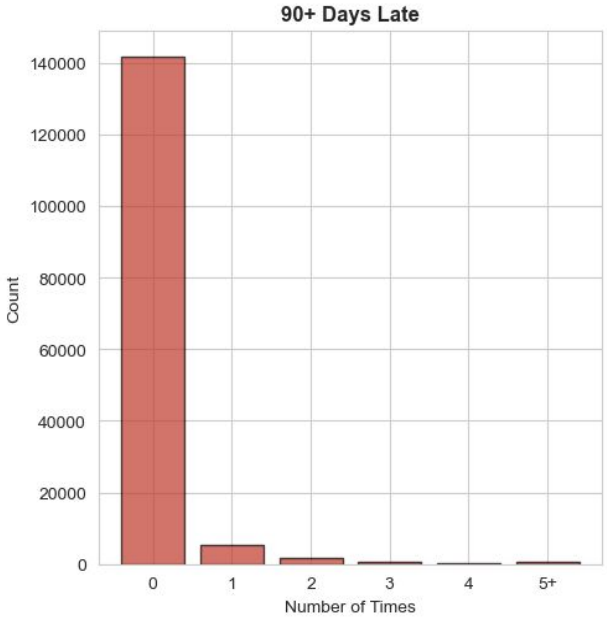


EDA

Late Payment Behavior

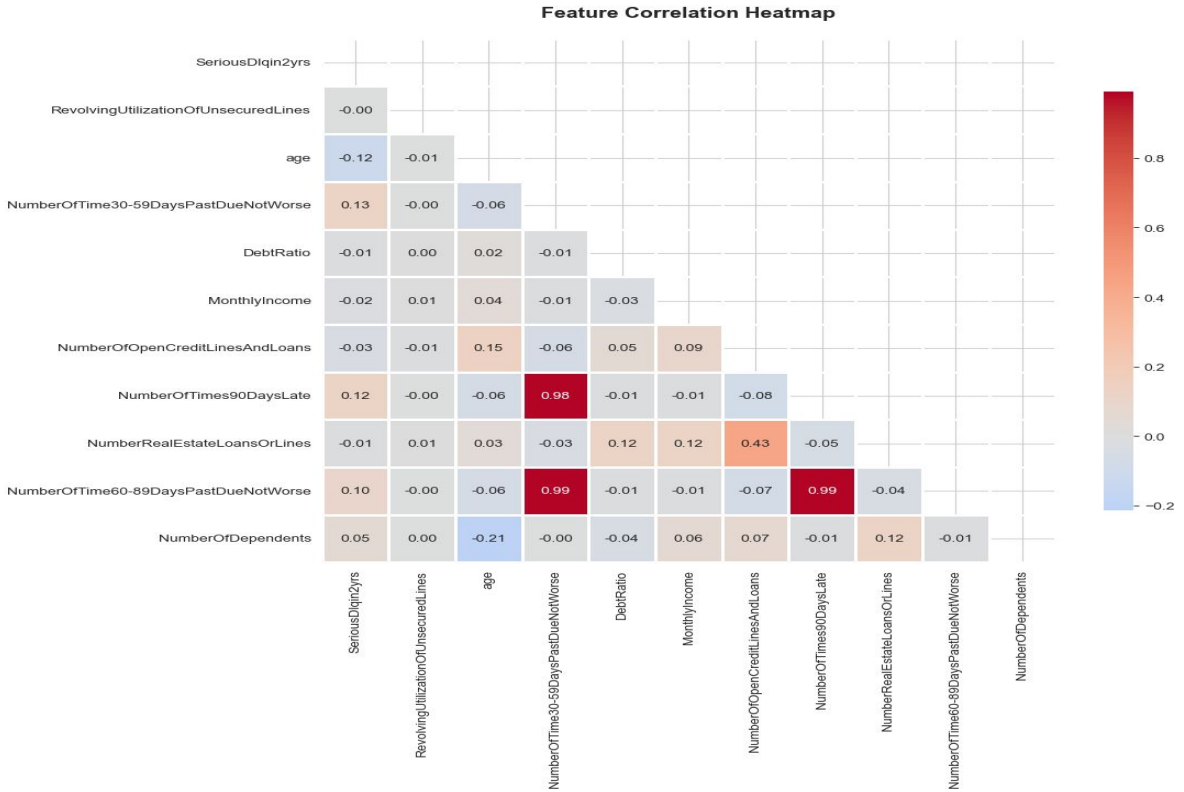






EDA

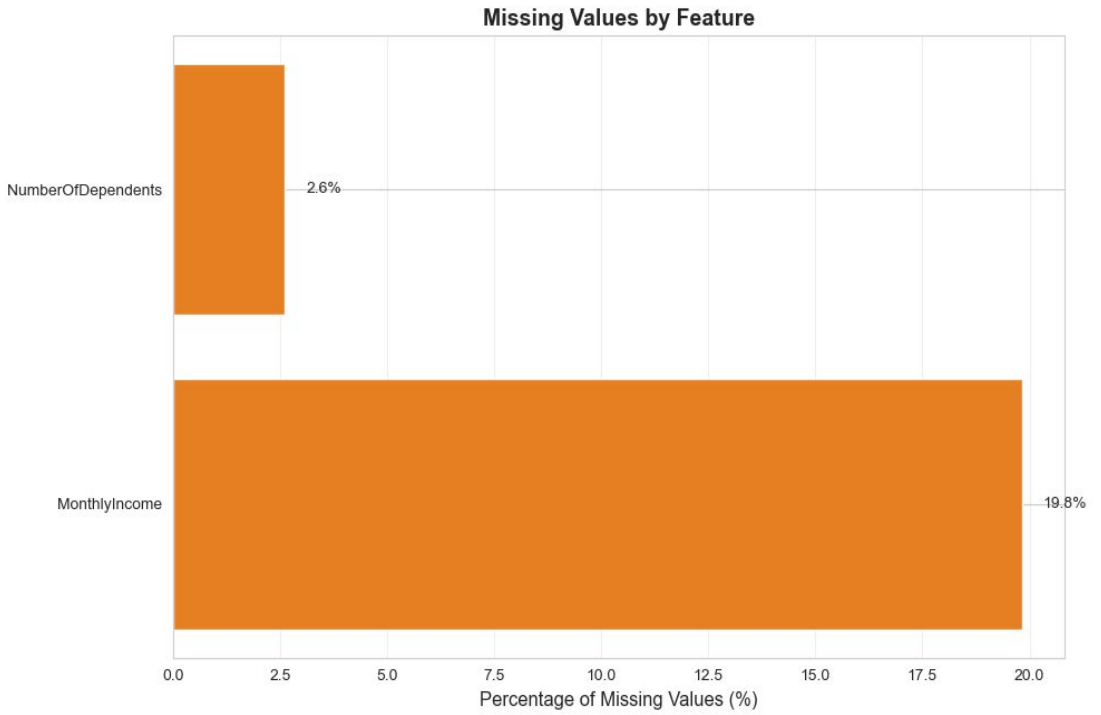
Feature Correlation Analysis



Data Preprocessing

Missing Value Analysis

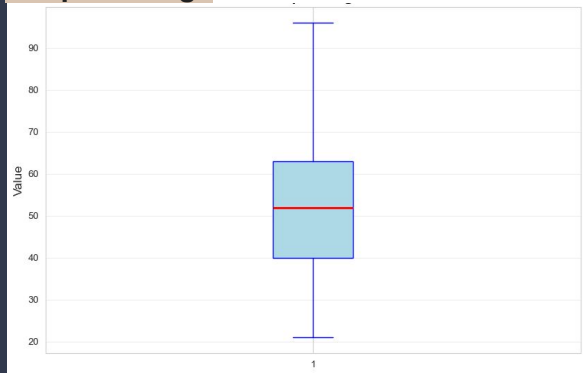
```
SeriousDlqin2yrs                         0
RevolvingUtilizationOfUnsecuredLines     0
age                                      0
NumberOfTime30-59DaysPastDueNotWorse     0
DebtRatio                                0
MonthlyIncome                        29731
NumberOfOpenCreditLinesAndLoans          0
NumberOfTimes90DaysLate                  0
NumberRealEstateLoansOrLines             0
NumberOfTime60-89DaysPastDueNotWorse     0
NumberOfDependents                    3924
dtype: int64
```



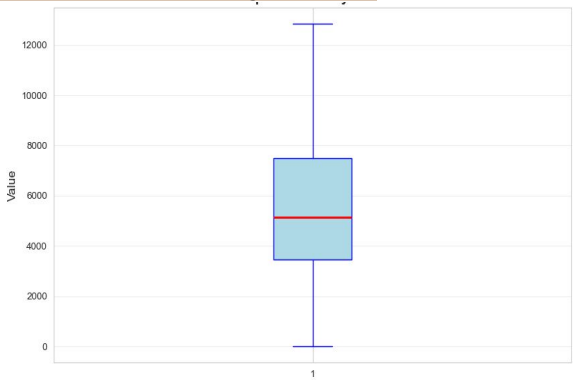
Data Preprocessing

Handling Outliers

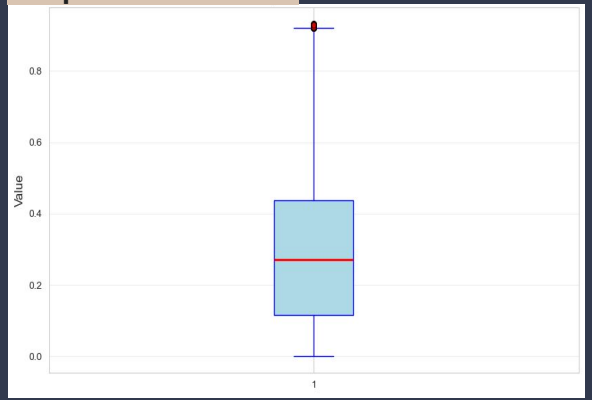
Boxplot of Age



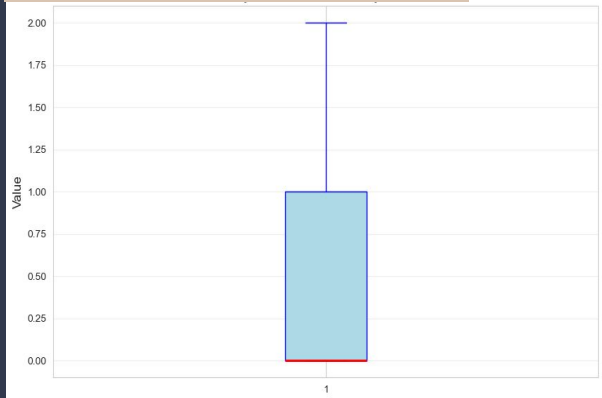
Boxplot of Monthly Income



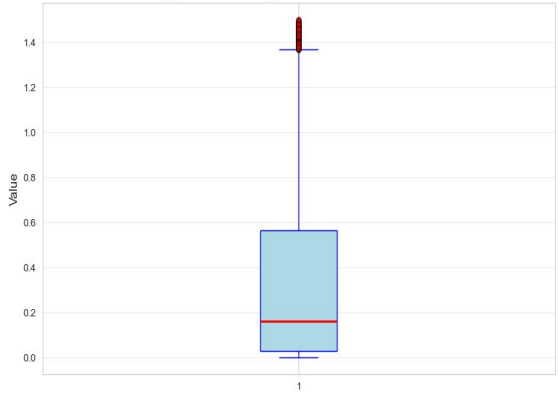
Boxplot of DebtRatio



Boxplot of NumberOfDependents



RevolvingUtilizationOfUnsecuredLines



Model and Evaluation

- ✔ Logistic Regression
- ✔ Random Forest Classifier
- ✔ Gradient Boosting Classifier
- ✔ MLP Classifier

```
pipes = {}

for acronym, model in models.items():
    if acronym in ['lr','mlpc']:
        pipes[acronym] = Pipeline([
        ('scaler', StandardScaler()),
        ('smote', SMOTE(random_state=random_seed)),
        ('model', model),
        ])

    else:
        pipes[acronym] = Pipeline([
            ('smote', SMOTE(random_state=random_seed)),
            ('model', model),
        ])

pipes
```

Hyperparameter Tuning

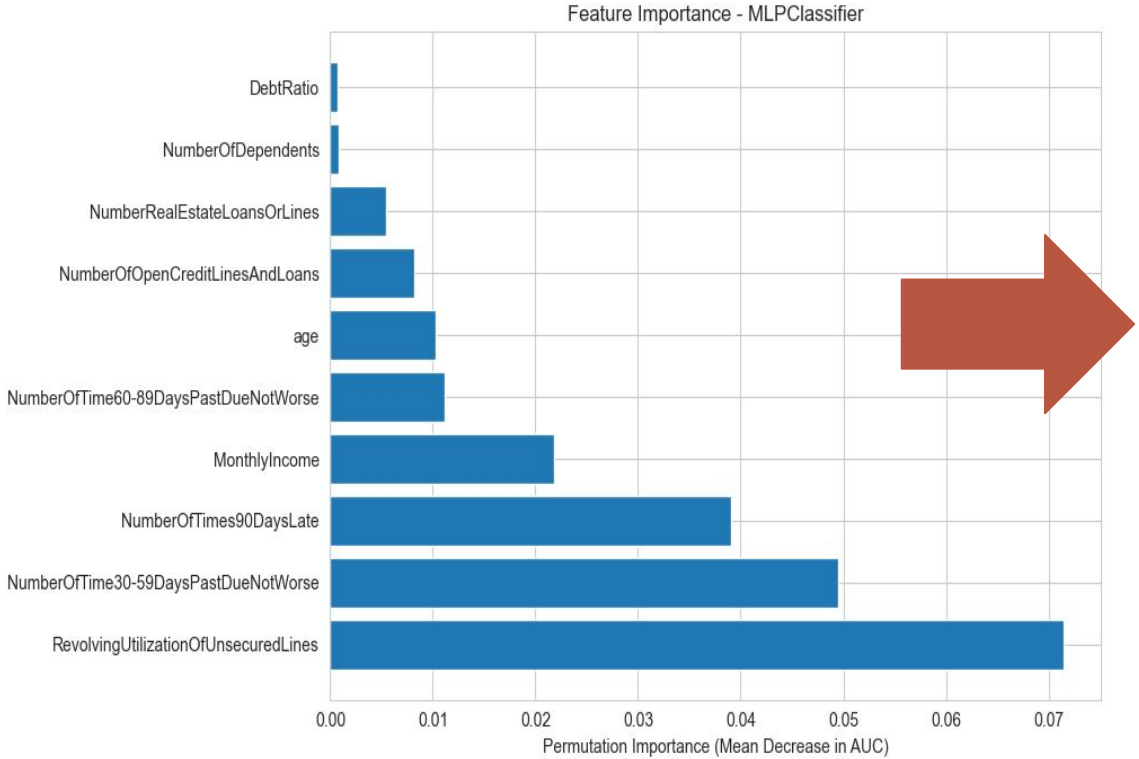
GridSearch CV

Logistic Regression	• model__C : [0.1, 1, 10], • model__penalty : ['l2'], • model__class_weight : [None, 'balanced']
Random Forest Classifier	• model__n_estimators : [200, 400], • model__max_depth : [5, 10, None], • model__min_samples_split : [2, 5], • model__class_weight : [None, 'balanced_subsample']
Gradient Boosting Classifier	• model__n_estimators : [100, 200], • model__learning_rate : [0.05, 0.1], • model__max_depth : [2, 3]
MLP Classifier	• model__alpha : alpha_grids, • model__learning_rate_init : lr_init_grids, • model__hidden_layer_sizes : hidden_layer_sizes_grids

Model Result

Model	best_cv_auc	test_auc
MLP Classifier	**0.851182**	**0.840149**
Random Forest Classifier	0.828728	0.821720
Gradient Boosting Classifier	0.826637	0.815682
Logistic Regression	0.800074	0.794796

Feature Importance




```
============================================================
REDUCED FEATURE MODEL PERFORMANCE
============================================================
Test AUC:               0.8190
Test Accuracy:          0.9366
Balanced Accuracy:      0.5753

Original Model AUC:     0.8495
AUC Difference:         -0.0305
============================================================

CLASSIFICATION REPORT:
------------------------------------------------------------
                precision    recall  f1-score   support

No Delinquency       0.94      0.99      0.97     18559
   Delinquency       0.62      0.16      0.25      1345

      accuracy                           0.94     19904
     macro avg       0.78      0.58      0.61     19904
  weighted avg       0.92      0.94      0.92     19904

CONFUSION MATRIX:
------------------------------------------------------------
[[18430   129]
...
True Negatives:  18430
False Positives: 129
False Negatives: 1133
True Positives:  212
```

Thank you